Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes

Sharad Goel‡, Andrei Broder†, Evgeniy Gabrilovich†, Bo Pang†

‡ Yahoo! Research, 111 West 40th Street, New York, NY 10018, USA
† Yahoo! Research, 4301 Great America Parkway, Santa Clara, CA 95054, USA

{goel, broder, gabr, bopang}@yahoo-inc.com

ABSTRACT

The success of "infinite-inventory" retailers such as Amazon.com and Netflix has been ascribed to a "long tail" phenomenon. To wit, while the majority of their inventory is not in high demand, in aggregate these "worst sellers," unavailable at limited-inventory competitors, generate a significant fraction of total revenue. The long tail phenomenon, however, is in principle consistent with two fundamentally different theories. The first, and more popular hypothesis, is that a majority of consumers consistently follow the crowds and only a minority have any interest in niche content; the second hypothesis is that everyone is a bit eccentric, consuming both popular and specialty products. Based on examining extensive data on user preferences for movies, music, Web search, and Web browsing, we find overwhelming support for the latter theory. However, the observed eccentricity is much less than what is predicted by a fully random model whereby every consumer makes his product choices independently and proportional to product popularity; so consumers do indeed exhibit at least some a priori propensity toward either the popular or the exotic.

Our findings thus suggest an additional factor in the success of infinite-inventory retailers, namely, that tail availability may boost head sales by offering consumers the convenience of "one-stop shopping" for both their mainstream and niche interests. This hypothesis is further supported by our theoretical analysis that presents a simple model in which shared inventory stores, such as Amazon Marketplace, gain a clear advantage by satisfying tail demand, helping to explain the emergence and increasing popularity of such retail arrangements. Hence, we believe that the return-on-investment (ROI) of niche products goes beyond direct revenue, extending to second-order gains associated with increased consumer satisfaction and repeat patronage. More generally, our findings call into question the conventional wisdom that specialty products only appeal to a minority of consumers.

WSDM'10, February 4–6, 2010, New York City, New York, USA.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Economics, Measurement

Keywords

Long tail, infinite inventory

1. INTRODUCTION

The explosion of electronic commerce has opened the door to so-called "infinite-inventory" retailers, such as Amazon.com, Netflix, and the iTunes Music Store, which offer an order of magnitude more items than their brick-and-mortar counterparts [2]. The resulting *long tail markets* [1, 2] have been found to exhibit two near-universal properties: (1) the vast majority of products are "misses," appealing to only a relatively small group of people; and (2) these "worst-sellers" in aggregate account for a sizable fraction of total consumption. For example, 30% of Amazon.com's sales and 25% of Netflix's sales are for items not available in the largest offline retail stores [2]. Based on these empirical observations, the success of online retailers has been largely attributed to the lucrative, and previously untapped, "tail markets."

The long tail phenomenon, however, is in principle consistent with two fundamentally different hypotheses. The first, and generally accepted theory, is that a majority of consumers prefer popular offerings while only a minority seek niche content; the second hypothesis is that everyone is a bit eccentric, consuming both popular and specialty products. These two theories, importantly, predict substantively different tradeoffs between inventory size and user satisfaction. In the former case, a small inventory of popular items would satisfy most people nearly all of the time, while in the latter, such an inventory would frustrate most people at least some of the time. Thus, differentiating between the two is key to developing sound business strategies.

To distinguish between these possible alternatives, we examine extensive data on user preferences for movies, music, Web search, and Web browsing. In all of these domains, we find overwhelming evidence that nearly everyone is at least a bit eccentric. Our findings suggest an additional factor for the success of infinite-inventory retailers: tail availability may boost head sales by offering consumers the convenience of "one-stop shopping" for both their mainstream and niche

interests. Hence, even small increases in direct revenue from niche products may be associated with much larger second-order gains due to increased overall consumer satisfaction and resulting repeat patronage. More generally, our work highlights the diversity of individual tastes, and calls into question the conventional view that niche products appeal only to a minority of consumers.

The remainder of our paper is organized as follows. In Section 2 we review related work. Section 3 describes our data and presents the main empirical findings. We propose and analyze a theoretical model of consumer behavior in Section 4, and discuss how small differences in inventories lead to "winner-take-all" effects. We conclude in Section 5 by summarizing and discussing our results.

2. RELATED WORK

The "long tail" view was coined and popularized by Chris Anderson [1, 2] to describe consumers' demand for niche products in an age of infinite-inventory retailers. In particular, Anderson finds that a substantial fraction of revenue is generated from specialty items not available in traditional brick-and-mortar stores, and argues that the "future of business is selling less of more" [2]. The economics of long tail markets have been further analyzed by Brynjolfsson et al. [4, 5, 6], who provide a theoretical framework and empirical detail. They consider drivers that increase the collective share of niche products both on the supply-side (e.g., lower stocking and distribution costs) and the demand-side (e.g., improved recommendation and search tools). On the other hand, Elberse et al. [9, 10] have suggested that tail inventory is overrated. Noting that the number of DVD titles in the top 10% of weekly sales dropped by more than 50% from 2000 to 2005, they conclude the importance of best sellers has been growing, not diminishing, over time. And Tan et al. [17], after adjusting for increasing product variety, likewise find that demand for hits has been rising.

In contrast to past work, which primarily considers the volume of tail sales, we focus on consumer satisfaction and the resulting second-order effects of tail inventory. By focusing on the consumer, we shed light on—and largely refute—the perception that niche content appeals only to a minority of consumers. In part, this misconception may be traced to what Levine describes as the "emergence of a cultural hierarchy" in early twentieth century America that established a stark divide between "lowbrow" and "highbrow" entertainment [12]. Looking primarily at high-status individuals, Peterson et al. suggest a relatively recent "historical shift from highbrow snob to omnivore is taking place" [14, 15]. Although we do not explicitly address the cultural status of consumers' choices, our results are consistent with this view of "omnivorous" individuals.

Elberse [9], writing in Harvard Business Review, reaches qualitative conclusions similar to some of our observations. Specifically, she posits in part that: (1) "a large number of customers occasionally select obscure offerings;" and (2) "customers with a large capacity for content venture into the tail." We provide extensive empirical evidence to support and refine these statements, and analyze, both empirically and theoretically, the consequences of these results on business strategies. Elberse further argues that consumers appreciate obscure movies less than popular movies, and thus advises retailers to "resist the temptation to direct customers to the tail." While we find—in agreement with Elberse—that popular movies receive the highest user ratings, the opposite appears to be true with music: The highest average ratings on Yahoo! Music are given to the most obscure songs. Furthermore, even in the case of movies, we find that typical users regularly give high marks to tail inventory (cf. Section 3.3.2).

Many authors have examined Web search query distributions. Spink et al. [11, 16] studied query logs of the Excite search engine, and analyzed basic properties of this query stream. Later, Downey et al. [7, 8] juxtaposed rare and common queries with rare and common information goals, and described distinctions in user behavior observed for queries and goals of differing rarity. We believe, however, that the long tail phenomenon previously has not been explicitly addressed in the context of Web search.

3. EMPIRICAL ANALYSIS

3.1 Data Description

Our empirical results are based on an analysis of user behavior across five large datasets: (1) ratings on the movie rental service Netflix; (2) ratings on the music service provider Yahoo! Music; (3) queries on Yahoo! Search; (4) clicked search results on Yahoo! Search; and (5) Web browsing activity collected by the Nielsen Company. Summary statistics for these datasets are given in Table 1.

We examined nearly 100 million Netflix movie ratings from over 400,000 users, and over 700 million Yahoo! Music ratings from over two million users. Netflix ratings were collected between November 1999 and December 2005, and Yahoo! Music ratings were collected between 2002 and 2006. As we are primarily concerned with user-centric statistics, we excluded users for whom we have limited data. In particular, the Netflix dataset was trimmed to include only users who had rated at least 10 movies, and the Yahoo! Music dataset was trimmed to include only users that rated at least 20 songs.[1] Additionally, the music dataset was comprised only of songs that received at least 20 ratings. In both datasets, users rated items (i.e., movies and songs, respectively) on a five point scale, and the primary incentive for users to rate items was to receive personalized recommendations. In neither case was there a requirement that users have purchased or intend to purchase the items they rate. Although these rating records are distinct from purchase histories, we believe they provide a reasonable indication of user interests.

For Web search related data, we analyzed one month of Yahoo! search logs (September, 2008). Simple transformations (e.g., mild stemming) were applied to collate equivalent queries. Furthermore, URLs for clicked search results were truncated to only include domains; for example, `http://en.wikipedia.org/wiki/Long_tail` was normalized to `en.wikipedia.org`. Infrequent users—those who issued fewer than 10 queries, or clicked on fewer than 10 URLs in the month-long dataset—were excluded from our analysis. In total, we considered approximately 2.6 billion queries and 2.5 billion click events across nearly 60 million users.

Analysis of Web browsing behavior was based on complete activity logs for the approximately 100,000 users in the Nielsen MegaPanel for the month of March, 2009. Users in

[1]Trimming retains nearly all users in these datasets.

	Movies	Music	Search Queries	Clicked Search Results	Web Browsing
Items	17,770	702,896	512,323,034	20,301,327	2,012,617
Users	429,541	2,156,792	57,524,526	57,758,157	109,315
Observations	99,548,085	755,480,158	2,613,137,669	2,491,026,154	287,189,911

Table 1: Descriptive statistics for the five datasets analyzed. Observations correspond to ratings, queries, click events, and page views, as appropriate to the domain. Movie data obtained from Netflix; music, search queries and clicked search results data obtained from Yahoo!; and Web browsing data obtained from the Nielsen Company.

the panel study were weighted based on their demographic attributes to mimic a representative sample of the U.S. online population. As with the search data, Web domains were extracted from visited URLs; in aggregate, users in our sample visited over two million unique domains and registered nearly 300 million page views.

3.2 The Long Tail of Consumption

Consistent with past work on the long tail, we find in all five datasets that: (1) a relatively small number of items account for a disproportionately large fraction of total consumption; and (2) the tail, in aggregate, is nevertheless relatively heavy. We define the *popularity* of an item (e.g., a movie, a song, or a URL) to be the fraction of total consumption fulfilled by that item. For example, the popularity of a given movie is defined to be the total number of times it was rated divided by the total number of movie ratings. Ranking items by their popularity (with lower ranks corresponding to greater popularity), we consider inventories of the k most popular items. For movies and music, Figures 1(a) and 1(b) plot cumulative popularity as a function of the inventory size k (e.g., the fraction of total ratings that are for movies in the top-k inventory). In particular, while the 100 most popular movies account for nearly 15% of consumption, the 3,000 most popular movies—the number available in a typical brick-and-mortar DVD retailer—still leave 13% of consumption unmet. Similarly, though the 1,000 most popular songs satisfy 13% of demand, the 50,000 most popular—the number available at a large, physical music retailer such as Wal-Mart—leave 34% of consumption unfulfilled.

In the case of Web search and Web browsing, though there is no analog to a physical retailer, we still observe these same two characteristics. A mere ten Web sites account for over 15% of page views, while the top 10,000 still leave over 20% of consumption unaccounted for. This relationship for all five domains is illustrated in Figure 1(c), with a log-log plot of popularity vs. normalized rank (i.e., rank divided by the total number of unique items in the given domain).

3.3 Individual Eccentricity

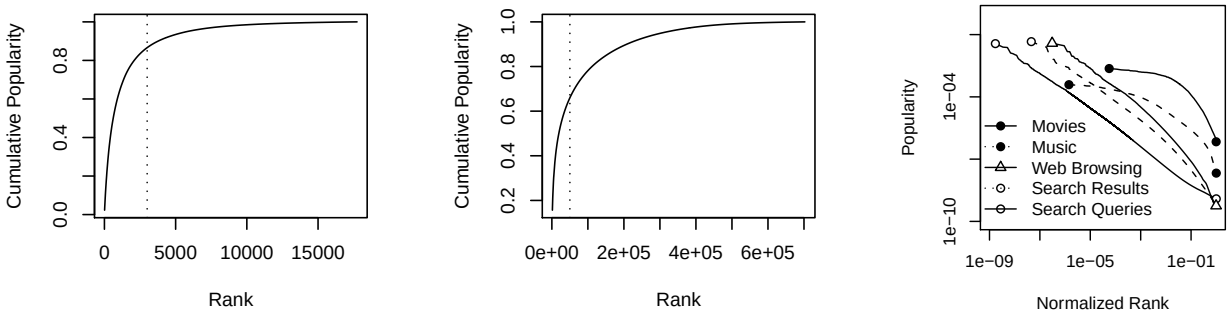
3.3.1 Tail Interests and Customer Satisfaction

Next we begin to address our central question: To what extent do individuals have niche interests? In particular, are most people satisfied by a relatively small inventory of high demand items—as suggested by conventional wisdom—or do people have more idiosyncratic tastes?

For any given inventory of the k most popular items, we say that a user is *p-percent satisfied* if at least p percent of the items they consume (e.g., rate, click, view, etc.) are contained in the inventory. In the following we focus primarily on 100% and 90% satisfaction. By this definition of satisfaction, we find that only 11% of Netflix users and only 5% of Yahoo! Music users are 100% satisfied by inventories comparable in size to those of large physical retailers (i.e., $k = 3,000$ and $k = 50,000$, respectively). In other words, almost none of the users in these movie and music datasets exclusively rate items likely to be found in large, offline retailers. Moreover, only 63% of Netflix users and only 32% of Yahoo! Music users are at least 90% satisfied by these limited—though seemingly large—traditional inventories. That is, one in ten times, fairly typical movie and music consumers in our datasets would likely be frustrated by the inventories of brick-and-mortar sellers. Figure 2 displays the complete 100% and 90% satisfaction curves as a function of inventory size. While 100% satisfaction is approximately linear in inventory size, 90% satisfaction is concave, quickly increasing before leveling off.

This pattern of relatively eclectic, and hence hard to satisfy, user interests suggests the value of tail inventory extends beyond direct revenue, contributing to second-order benefits such as repeat patronage associated with increased customer satisfaction. For an inventory of the k most popular items, we compare: (a) cumulative popularity (i.e., the fraction of demand fulfilled by the inventory); and (b) the fraction of users 100% (or 90%) satisfied by the inventory. Figure 3 plots the relationship between these two measures. In all of the five datasets, we see that small increases in cumulative popularity are associated with disproportionately large increases in satisfaction. For example, by moving from an inventory of 3000 movies to 3500 movies, cumulative popularity increases 2% (from 87% to 89%) while 90% satisfaction increases 7% (from 63% to 70%). In other words, movies that in and of themselves account for only 2% of demand could potentially grow the overall customer base 7% by attracting newly satisfied consumers. Consequently, to the extent that increased satisfaction attracts both the mainstream and niche consumption of new customers, direct revenue calculations undervalue the tail.

While the revenue effects discussed above are particularly salient in the case of movie and music sales, an analogous interpretation holds for Web search. Providing high quality search results for a rare query class often requires considerable effort and expense. Before investing in such a project, it is hence natural to ask how many such queries are issued. Our results, however, suggest the importance of an additional question: How many *users* issue such queries? Examining the search query dataset, we find that fulfilling an additional 1% of consumption in the tail, by moving from 95% to 96% consumption fulfilled, results in a 6% increase in 90% satisfaction—from 80% to 86%. Supporting rare queries, that is, can disproportionately increase overall user satisfaction.



(a) The long tail of Netflix. The dotted vertical line at 3,000 indicates the typical inventory size of a large brick-and-mortar retailer.

(b) The long tail of Yahoo! Music. The dotted vertical line at 50,000 indicates the typical inventory size of a large brick-and-mortar retailer.

(c) Long tails in movies, music, Web browsing, clicked search results, and search queries.

Figure 1: The long tails of music, movies, search queries, clicked search results, and Web browsing.

3.3.2 Double Jeopardy

In his influential book *Formal Theories of Mass Behavior* [13], William McPhee predicted that the more obscure an item, the less likely it was to be appreciated by those who came across it. Hence, he characterized niche products to be in a state of *double jeopardy*—first, they were not generally known, and second, they were not generally liked by those who did know of them. Empirical support for this position is found both in Elberse's analysis of the Quickflix DVD rental data [9], and also in our own examination of Netflix (Figure 4(a)): in both movie datasets, average consumer ratings increase with popularity. Interestingly, however, a very different pattern emerges in the music dataset. Both the most popular and the least popular songs receive the highest ratings, with a dip in the middle of the inventory. In fact, the most obscure songs receive slightly higher average ratings than the most popular ones (Figure 4(b)).

Citing the increase of movie ratings with popularity, Elberse [9] suggests that the value of the tail has been overstated since users are disproportionately dissatisfied with niche inventory. To investigate this claim, we restricted our datasets and for each user only considered movies and songs that they rated highly (i.e., gave at least 4 out of 5 stars); the corresponding satisfaction curve for Netflix movies is plotted in Figure 4(c). In particular, we find that 85% of Netflix users and 91% of Yahoo! music users rated highly a movie or song not likely to be found in a large, physical retailer (i.e., $k = 3,000$ and $k = 50,000$, respectively). Moreover, for 32% of Netflix users and for 56% of Yahoo! Music users, at least 10% of the items they rated highly were in the tail. Consequently, it seems hard to dismiss what appears to be widespread interest and appreciation for tail content.

3.3.3 A "Null Hypothesis" Model of Consumer Preferences

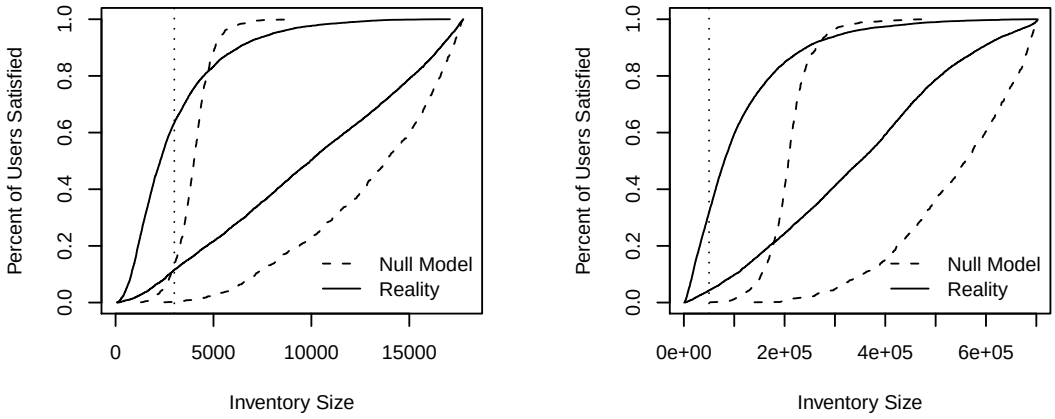
The results above—which show that even typical users have a relatively high demand for tail items—suggest the following simple null model of user behavior. First, each user randomly decides how many items to consume (i.e., rate, view, click, etc.), adhering to the empirically observed, domain-specific distribution of individual user consumption. Users then randomly select items to consume proportional to item popularity, with the selection being done without replacement (i.e., a user cannot select the same item more than once). Note that this model approximately preserves both the empirical distribution of user consumption and the empirical distribution of item popularity, but otherwise disregards any personal preferences for head or tail items, and hence represents the "null hypothesis" for the existence of such propensities.

Figure 2 displays the 100% and 90% satisfaction curves for the movie and music null models as a function of inventory size, together with the empirically observed ("reality") curves. By construction these models capture the fact that typical users regularly consume items not available at limited-inventory retailers. However, the null models predict that users are *much harder* to satisfy than what we observe in reality. In particular, only 14% of users in the movie model are at least 90% satisfied by brick-and-mortar sized inventories (compared to 63% in the data); and approximately none of the users in the music model are at least 90% satisfied by offline inventories (compared to 32% in the data). This indicates that in reality there is a sizable fraction of users whose preference for head over tail content extends beyond the relative popularity of the head over the tail, and the null hypothesis can be rejected. That is, although nearly all users consume tail content at least part of the time, it appears that some users draw disproportionately from the head while others draw disproportionately from the tail.

3.3.4 Individual Variation in Taste

To further investigate this variance in individual preferences, for each user we define his *eccentricity*—the median rank of items he has consumed, where items, as before, are ordered by popularity. In particular, higher eccentricity corresponds to on average consuming less popular items.

Figure 5 shows the distribution of eccentricity among users for movies, music, and Web browsing. In all three domains, there is significant variation between individuals, evidenced by relatively wide interquartile ranges: $[226 - 683]$,



(a) Netflix. The dotted vertical line at 3,000 indicates the typical inventory size of a large brick-and-mortar retailer. The two upper curves are 90% satisfaction, and the two lower curves are 100% satisfaction.

(b) Yahoo! Music. The dotted vertical line at 50,000 indicates the typical inventory size of a large brick-and-mortar retailer. The two upper curves are 90% satisfaction, and the two lower curves are 100% satisfaction.

Figure 2: 90% and 100% user satisfaction curves for movies and music as a function of inventory size. The solid lines correspond to empirically observed data, and the dashed lines represent a null model where users select items proportional to popularity.

$[2774 - 16890]$ and $[138 - 2156]$, respectively. Furthermore, analogous to the results of Section 3.3.3, user eccentricity is considerably larger under the movie and music null models than is seen in the data. Interestingly, in the case of Web browsing, typical eccentricity under the null model is comparable to what is empirically observed; the significant difference, however, is that the empirical eccentricity distribution has a much heavier tail.[2]

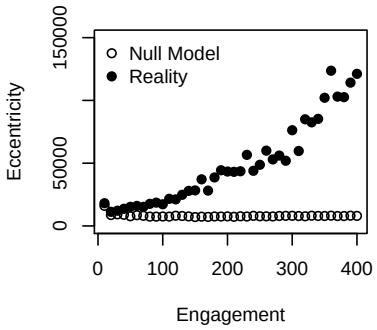


Figure 7: Average eccentricity vs. engagement for Web browsing, where we consider only unique sites.

Finally, we examine the relationship between eccentricity and *engagement*—the number of items an individual consumes. One might reasonably suspect that those who watch more movies, listen to more songs, or view more Web pages, are also more eccentric. Nevertheless, we find only weak correlations between engagement and eccentricity for movies (0.2), music (0.1), and Web browsing (-0.1), where the correlation is actually negative in the last case. Hence, at the level of individuals, engagement is not a strong predictor of eccentricity.

Individual variance, however, masks a pronounced effect of engagement on mean, group-level eccentricity. For example, those who rate approximately three thousand songs on Yahoo! Music are, on average, more than twice as eccentric (35,310) as those who rate approximately five hundred songs (16,821). Figure 6 shows this effect of engagement on average eccentricity, where users are binned on the x-axis according to their level of engagement, and the mean eccentricity of users in each bucket is given on the y-axis. In the case of movies and music, mean eccentricity increases linearly with engagement. Moreover, relative to the null model of Section 3.3.3, light users draw disproportionately from the head, and heavy users from the tail.

For Web browsing, however, where engagement corresponds to page views, mean eccentricity is approximately independent of engagement. Considering page views to be a proxy for time, this indicates that, on average, light and heavy Internet users spend their time on Web sites of comparable popularity. However, if instead of page views we consider unique URLs visited, we see that mean eccentricity increases with engagement, analogous to movies and music (Figure 7).[3]

[2]In the Web browsing null models, users select items (i.e., URLs) with replacement, since users may view—and regularly do view—the same Web page more than once.

[3]Specifically, the popularity of a site is now defined to be the number of users who have visited it at least once, a user's engagement is the number of unique URLs she has visited, and her eccentricity is the median rank of these unique URLs.

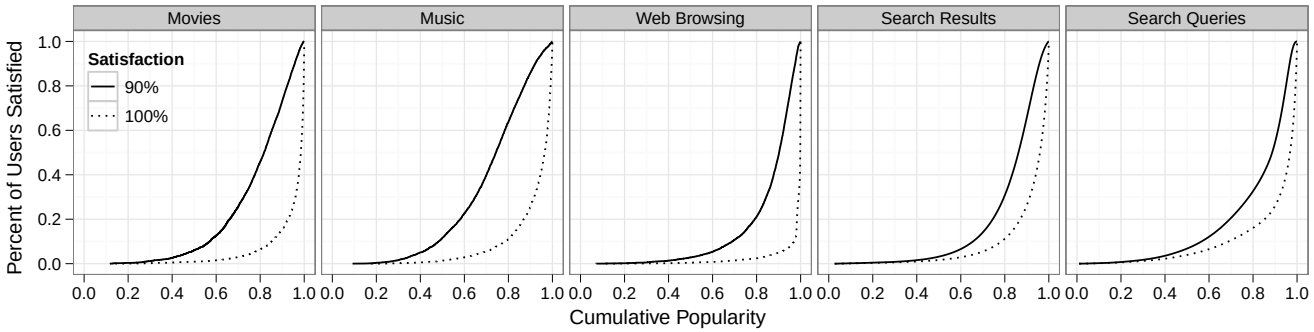
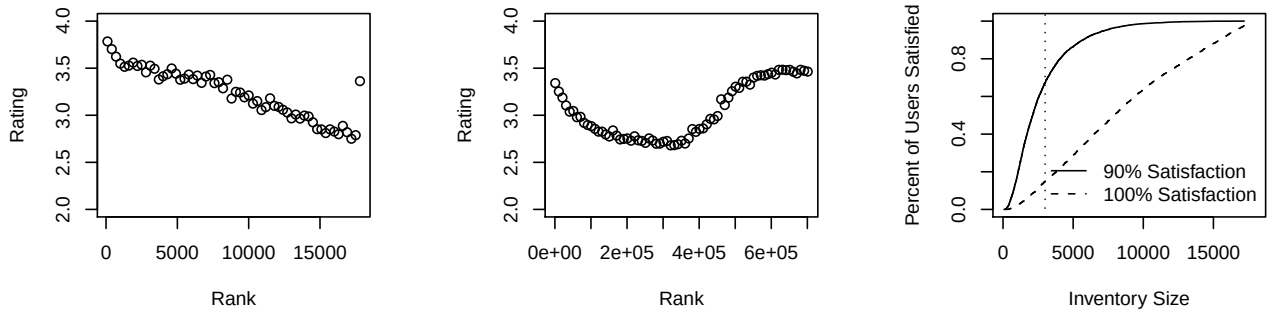


Figure 3: 90% and 100% satisfaction as a function of cumulative popularity for inventories of the top k items. In all five domains, small increases in cumulative item popularity are associated with disproportionately large increases in satisfaction.



(a) Average movie ratings as a function of popularity, where movies are binned into buckets of size 100.

(b) Average music ratings as a function of popularity, where songs are binned into buckets of size 1000.

(c) Satisfaction curves on Netflix when we restrict to highly rated movies.

Figure 4: An analysis of user-generated ratings for movies and music.

4. THEORETICAL ANALYSIS

We start by analyzing the shape of user satisfaction curves under a simple model of behavior. We then show, more generally, how small differences in inventory between competing retailers may lead to winner-take-all outcomes.

Consider a continuous analog of the null model in Section 3.3.3, where users independently select items proportional to popularity. For $\alpha > 1$, suppose the item popularity distribution is described by a power law with exponent α supported on $[1, \infty)$. That is, the cumulative distribution function (CDF) of popularity is given by $F_\alpha(x) = 1 - 1/x^{\alpha-1}$, and its corresponding density is $f_\alpha(x) = (\alpha-1)/x^\alpha$. In this case, Theorem 1 describes the shape of the 100% user satisfaction curve as a function of inventory size.

THEOREM 1. *Consider the selection model described above. For $k \geq 1$ an integer, suppose $X_1, \ldots, X_k$ are independent draws from the distribution F_α and let $M = \max(X_1, \ldots, X_k)$. Denote the CDF of M by $G_{k,\alpha}(x)$. Then $G_{k,\alpha}(x)$ has an inflection point at*

$$x^* = \left(1 + \frac{(k-1)(\alpha-1)}{\alpha}\right)^{1/(\alpha-1)}$$

and $G_{k,\alpha}(x)$ is convex for $x < x^*$ and concave for $x > x^*$. Furthermore,

$$G_{k,\alpha}(x^*) = \left(1 - \frac{1}{1 + \frac{(k-1)(\alpha-1)}{\alpha}}\right)^k \longrightarrow \exp\left(\frac{\alpha}{1-\alpha}\right)$$

where the limit is taken as $k \to \infty$.

PROOF. Begin by observing that the CDF of the max M is given by

$$G_{k,\alpha}(x) = [F_\alpha(x)]^k = \left[1 - \frac{1}{x^{\alpha-1}}\right]^k$$

for $x \geq 1$ and $G_{k,\alpha}(x) = 0$ otherwise. Taking derivatives, we have

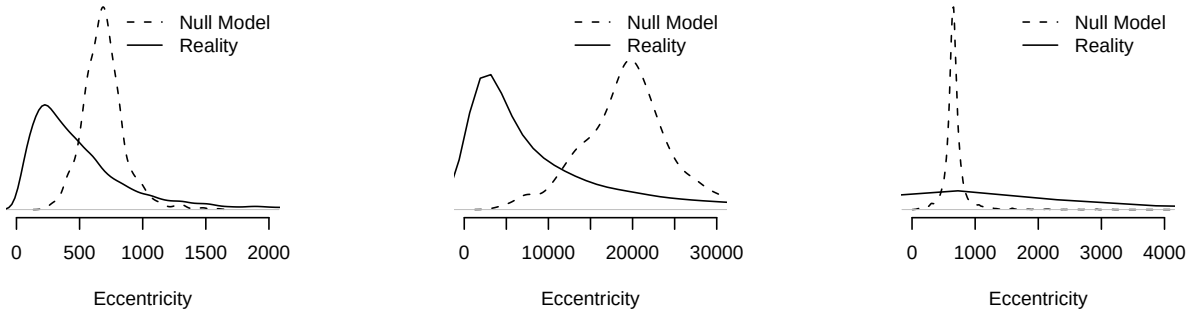
$$\frac{d}{dx} G_{k,\alpha}(x) = k(\alpha-1)\left[1 - \frac{1}{x^{\alpha-1}}\right]^{k-1} x^{-\alpha}$$

and

$$\frac{d^2}{dx^2} G_{k,\alpha}(x) = k(\alpha-1)\left(1 - \frac{1}{x^{\alpha-1}}\right)^{k-2} x^{-2\alpha} \times \left[(k-1)(\alpha-1) + \alpha - \alpha x^{\alpha-1}\right].$$

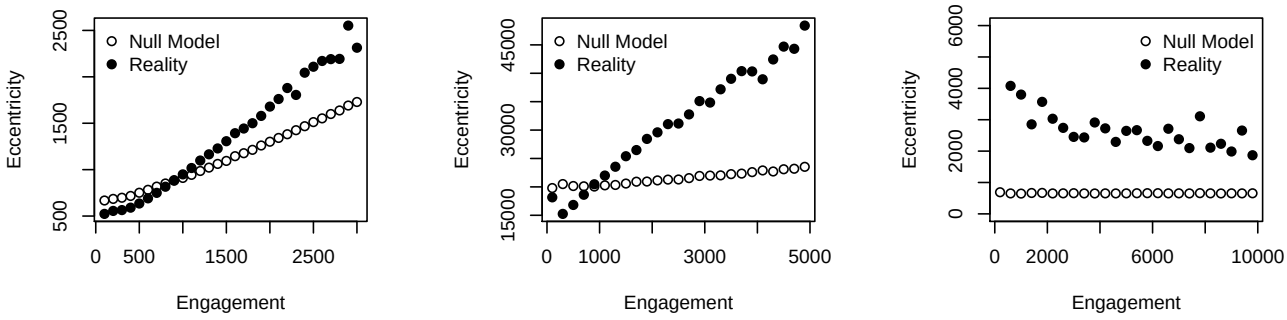
The first four terms are positive for $x > 1$, and so the unique

In the null model, selection is done without replacement.



(a) Netflix (b) Yahoo! Music (c) Web Browsing (Nielsen)

Figure 5: Distribution of user eccentricity for movies, music, and Web browsing, where eccentricity is the median rank of consumed items. In the null model, users select items proportional to item popularity.



(a) Netflix (b) Yahoo! Music (c) Web Browsing (Nielsen)

Figure 6: Average user eccentricity vs. engagement for movies, music, and Web browsing, where eccentricity is the median rank of consumed items and engagement corresponds to the number of ratings/page views. In the null model, users select items proportional to item popularity.

inflection point is given by the last term. That is,

$$x^* = \left(1 + \frac{(k-1)(\alpha-1)}{\alpha}\right)^{1/(\alpha-1)}.$$

Furthermore, it is clear that the second derivative is positive for $1 < x < x^*$ and negative for $x^* < x < \infty$, proving the concavity statements. Finally, evaluating $G_{k,\alpha}(x)$ at x^* yields

$$G_{k,\alpha}(x^*) = \left(1 - \frac{1}{1 + \frac{(k-1)(\alpha-1)}{\alpha}}\right)^k.$$

Taking the logarithm, together with a simple application of l'Hôpital's rule, shows that

$$\lim_{k\to\infty} G_{k,\alpha}(x^*) = \exp\left(\frac{\alpha}{1-\alpha}\right).$$

□

Intuitively, $X_1, \ldots, X_k$ in Theorem 1 indicate the ranks of k items selected by the consumer. Thus, $G_{k,\alpha}(x)$ can be interpreted as the probability the consumer would find all k items in an inventory of size x (i.e., that he would be 100% satisfied). What the theorem then shows is that satisfaction at first rapidly increases as a function of inventory size (i.e., is convex) before eventually leveling off.

Next we discuss how small differences in inventory can lead to large revenue disparities. Suppose there is a universe of items $\Omega = \{x_1, x_2, \ldots, x_N\}$, and two retailers A and B with respective inventories $I_A, I_B \subseteq \Omega$. We consider two descriptions of customer behavior, the *independent model* and the *sticky model*. In both cases, we assume there is a positive distribution Q such that customers select items $x_i \in \Omega$ with probability $Q(x_i)$.

- **Independent Model.** The customer first draws an item $x_i \in \Omega$ according to the distribution Q, and then selects one of the retailers A or B randomly with equal probability. If the selected retailer carries the selected item, she buys it; otherwise she checks for the item at the competing retailer, purchasing x_i there if it is available.

- **Sticky Model.** The customer first draws an item $x_i \in \Omega$ according to the distribution Q. However, instead of randomly selecting a retailer, she first searches the retailer from whom she most recently purchased an item. As before, she buys the item from that retailer if it is available, otherwise she attempts to purchase the item from the competing retailer. Note that at the start of the process, the customer selects an initial preferred retailer where she begins her searches up until making her first purchase.

Theorem 2 derives the long term fraction of sales completed by each retailer under these two models of consumer behavior.

THEOREM 2. *Consider the independent and sticky models of customer behavior described above. For a given customer u, set $Y_i = 1$ if the i^{th} item u attempts to purchase was eventually obtained at retailer A, and set $Y_i = 0$ otherwise. Under the independent model*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} Y_i = Q(I_A \cap \overline{I_B}) + \frac{1}{2}Q(I_A \cap I_B) \quad a.s.$$

Under the sticky model:

1. *If $I_A = I_B$ and u's initial preferred retailer is A, then*
$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} Y_i = Q(I_A) \quad a.s.$$
2. *If $I_A = I_B$ and u's initial preferred retailer is B, then $Y_i = 0$ for all i.*
3. *If $I_A \neq I_B$, then regardless of u's initial preferred retailer*
$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} Y_i = Q(I_A)Q(I_A|I_A\Delta I_B) + Q(I_A \cap \overline{I_B})Q(I_B|I_A\Delta I_B) \quad a.s.$$
where $I_A\Delta I_B = (I_A \cup I_B) \setminus (I_A \cap I_B)$ indicates the symmetric difference of the two inventory sets.

PROOF. We first analyze the independent model. Let x_i be the i^{th} item selected by u, and note that the indicator variables Y_i are independent and identically distributed. Then $Y_i = 1$ if either: (1) $x_i \in I_A \cap \overline{I_B}$ (i.e., x_i can be found at retailer A but not at retailer B); or (2) $x_i \in I_A \cap I_B$ and u first searched for x_i at A. Consequently,

$$P(Y_i = 1) = Q(I_A \cap \overline{I_B}) + \frac{1}{2}Q(I_A \cap I_B).$$

The result now follows from the strong law of large numbers.

Next we consider the sticky model with $I_A = I_B$. Then u will never switch retailers, since if she can't find what she is looking for at one retailer, than she will not find it at the other retailer either. Consequently, if she initially prefers B, she will never buy anything from A (i.e., $Y_i = 0$ for all i). On, the other hand, if she initially prefers A, the Y_i are independent and identically distributed with expectation $EY_i = Q(I_A)$. The result now again follows from the strong law of large numbers.

We analyze the case $I_A \neq I_B$ by a Markov chain on three states: $\{A_s, A_f, B\}$, where A_s indicates u just successfully purchased an item from retailer A, A_f indicates that u failed to purchase an item from A but will still attempt to purchase her next item from A first (i.e., B also failed to stock the item), and B indicates that u will attempt to purchase her next item from B first. The transition matrix K for this Markov chain is as follows:

	A_s	A_f	B
A_s	g	$1-(p+g)$	p
A_f	g	$1-(p+g)$	p
B	q	0	$1-q$

where

- $p = Q(\overline{I_A} \cap I_B)$ is the probability of not finding an item at A but finding it at B; hence, p is the probability of transitioning from A_s to B, and also the probability of transitioning from A_f to B.
- $q = Q(\overline{I_B} \cap I_A)$ is the probability of not finding an item at B but finding it at A; hence, q is the probability of transitioning from B to A_s.
- $g = Q(I_A)$ is the probability an item is carried by A; hence, g is the probability of transitioning from A_f to A_s, and also the probability of transitioning from A_s and A_s.

Note that since $I_A \neq I_B$, $p + q = Q(I_A\Delta I_B) > 0$.

We are interested in the long run average time the chain spends in state A_s. The stationary distribution for K is given by the unique row vector π such that $\pi K = \pi$ and such that the entries of π sum to 1. Finding π consequently reduces to computing the (right) null space of $K^T - I$, where I is the identity matrix. From an elementary computation, it follows that

$$\pi = \left(\frac{q(p+g)}{p+q}, \frac{q[1-(p+g)]}{p+q}, \frac{p}{p+q}\right).$$

Now,

$$\begin{aligned}\pi(A_s) &= \frac{q(p+g)}{p+q} \\ &= \frac{Q(\overline{I_B} \cap I_A)Q(\overline{I_A} \cap I_B)}{Q(I_A\Delta I_B)} + \frac{Q(\overline{I_B} \cap I_A)Q(I_A)}{Q(I_A\Delta I_B)} \\ &= Q(I_A)Q(I_A|I_A\Delta I_B) + Q(I_A \cap \overline{I_B})Q(I_B|I_A\Delta I_B).\end{aligned}$$

Finally, the result follows from the ergodic theorem for Markov chains (see e.g., Theorem 4.1 in [3]). □

When $I_A \neq I_B$, Theorem 2 shows that the fraction of demand ultimately fulfilled at retailer A is a weighted average of $Q(I_A)$ and $Q(I_A \cap \overline{I_B})$, where the weight on $Q(I_A)$ is given by $Q(I_A|I_A\Delta I_B)$. This result thus highlights the importance in the sticky model of stocking distinctive content (i.e., content only available at one retailer).

Moreover, unlike in the independent model, the sticky model can result in winner-take-all dynamics. Specifically, in the independent model, since

$$\begin{aligned}Q(I_A \cap \overline{I_B}) + \frac{1}{2}Q(I_A \cap I_B) &= Q(I_A) - \frac{1}{2}Q(I_A \cap I_B) \\ &\geq \frac{1}{2}Q(I_A)\end{aligned}$$

A will convert at least $Q(A)/2$ fraction of demand to sales, regardless of B's inventory. In the sticky model, however, if

$I_A \subsetneq I_B$ (i.e., if B has everything A has plus some more), then $Q(I_A|I_A \Delta I_B) = 0$, and A will consequently complete a vanishing fraction of sales since the customer will never have reason to switch from B to A.

As seen above, repeat business from satisfied customers can substantially outweigh revenue from any single purchase. This suggests significant gains are possible from so-called *shared inventory* business arrangements. As before, we suppose there are two retailers A and B with inventories $I_A, I_B \subseteq \Omega$. However, we further suppose that A partners with another merchant C, selling items on C's behalf when A does not itself stock an item. This model approximates Amazon Marketplace and eBay's Half.com, the key difference being that in our model A receives no direct revenue from selling C's items. Thus, the shared inventory model isolates the second-order benefits of hosting another merchant's inventory.

- **Shared Inventory Model.** As with the *sticky model*, the customer first selects an item $x_i \in \Omega$ according to Q, and then starts his search at the retailer A or B from whom he most recently purchased an item, searching the other retailer next if he cannot find what he is looking for. When the customer searches A, however, A makes available both its own inventory and the inventory of C. If both A and C carry the item, A sells its own copy. If C carries the item but A does not, then A sells the item on C's behalf, taking no direct profit from the sale. In this latter case the customer still begins his next search at retailer A.

THEOREM 3. *Consider the shared inventory model above. For a given customer u, set $Y_i = 1$ if the i^{th} item u attempts to purchase was eventually obtained at retailer A from A's own inventory, and set $Y_i = 0$ otherwise; set $Z_i = 1$ if this i^{th} item was eventually obtained at retailer A from C's inventory, and set $Z_i = 0$ otherwise. If $I_A \cup I_C \neq I_B$ then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} Y_i = Q(I_A)Q(I_A \cup I_C|(I_A \cup I_C)\Delta I_B) + Q(I_A \cap \overline{I_B})Q(I_B|(I_A \cup I_C)\Delta I_B) \quad a.s.$$

and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} Z_i = Q(I_C \cap \overline{I_A})Q(I_A \cup I_C|(I_A \cup I_C)\Delta I_B) + Q(I_C \cap \overline{I_A \cup I_B})Q(I_B|(I_A \cup I_C)\Delta I_B) \quad a.s.$$

regardless of u's initial preference.

PROOF. As with Theorem 2, we analyze a Markov chain on three states: $\{A_s, A_f, B\}$, where A_s indicates u successfully purchased an item from retailer A from A's own inventory, A_f indicates that u failed to purchase an item from A's inventory but will still attempt to purchase her next item from A first (i.e., either A sold her the item from C's inventory on C's behalf, or the item was not available from A, B or C), and B indicates that u will attempt to purchase her next item from B first. The transition matrix K for this Markov chain is as follows:

	A_s	A_f	B
A_s	g	$1-(p+g)$	p
A_f	g	$1-(p+g)$	p
B	q	r	$1-(q+r)$

where

- $p = Q(\overline{I_A \cup I_C} \cap I_B)$ is the probability of not finding an item in A's or C's inventory, but finding it at B; hence, p is the probability of transitioning from A_s to B, and also the probability of transitioning from A_f to B.
- $q = Q(\overline{I_B} \cap I_A)$ is the probability of not finding an item at B but finding it in$A$'s inventory; hence, q is the probability of transitioning from B to A_s.
- $g = Q(I_A)$ is the probability an item is in A's inventory; hence, g is the probability of transitioning from A_f to A_s, and also the probability of transitioning from A_s and A_s.
- $r = Q(\overline{I_A \cup I_B} \cap I_C)$ is the probability of not finding an item in A's or B's inventory, but finding it in C's inventory; hence r is the probability of transitioning from B to A_f.

Observe that $r + q = Q((I_A \cup I_C) \cap \overline{I_B})$, and hence $r + p + q = Q((I_A \cup I_C)\Delta I_B)$. In particular, since $I_A \cup I_C \neq I_B$, $r + p + q > 0$. An elementary computation shows that the stationary distribution for the Markov chain K is

$$\pi = \left(\frac{gr+qp+gq}{r+q+p}, \frac{r+q-(gr+qp+gq)}{r+q+p}, \frac{p}{r+q+p}\right).$$

Now,

$$\begin{aligned}
\pi(A_s) &= \frac{g(r+q)}{r+q+p} + \frac{qp}{r+q+p} \\
&= Q(I_A)\frac{Q((I_A \cup I_C) \cap \overline{I_B})}{Q((I_A \cup I_C)\Delta I_B)} \\
&\quad + Q(I_A \cap \overline{I_B})\frac{Q(\overline{I_A \cup I_C} \cap I_B)}{Q((I_A \cup I_C)\Delta I_B)} \\
&= Q(I_A)Q(I_A \cup I_C|(I_A \cup I_C)\Delta I_B) \\
&\quad + Q(I_A \cap \overline{I_B})Q(I_B|(I_A \cup I_C)\Delta I_B).
\end{aligned}$$

The result follows for $\lim_{n\to\infty} 1/n \sum_{i=1}^{n} Y_i$ now follows from the ergodic theorem for Markov chains [3].

To analyze $\lim_{n\to\infty} 1/n \sum_{i=1}^{n} Z_i$, first note that without loss of generality we may set $I_{\tilde{C}} = I_C \cap \overline{I_A}$ and analyze the behavior of $A, B, \tilde{C}$, since an item is purchased from C's inventory only if it is not in A's inventory. Now, since A and $\tilde{C}'s$ inventory is disjoint, we can interchange their roles (i.e., it is equivalent to first search $\tilde{C}$'s inventory, and then A's inventory). Consequently, we can use the result for $\lim_{n\to\infty} 1/n \sum_{i=1}^{n} Y_i$, replacing I_A with $I_C \cap \overline{I_A}$ and I_C with I_A. □

As in the sticky model, the fraction of demand ultimately fulfilled at retailer A under the shared inventory model is a weighted average of $Q(I_A)$ and $Q(I_A \cap \overline{I_B})$. However, the weight on $Q(I_A)$ increases from $Q(I_A|I_A\Delta I_B)$ under the sticky model to $Q(I_A \cup I_C|(I_A \cup I_C)\Delta I_B)$ under the shared inventory model. In particular, even though A does not receive any direct revenue from C, A nevertheless benefits from increased sales of its own merchandise.

5. DISCUSSION

Looking at extensive data on user preferences for movies, music, Web search, and Web browsing, we find overwhelming evidence that the vast majority of users are a little bit eccentric, consuming niche products at least some of the time. These results largely refute the conventional wisdom

that specialty products appeal only to a minority of consumers, and suggest that the benefit of tail inventory extends beyond direct revenue to second-order gains associated with increased consumer satisfaction and repeat patronage. Namely, as formalized by our sticky model of consumer behavior, tail inventory may boost head sales by providing users a convenient one-stop shop for both their mainstream and niche interests. Moreover, our analysis provides theoretical support for shared inventory business models such as Amazon Marketplace and Half.com.

Given the observed user eccentricity, one reasonable hypothesis is that users consume content proportional to popularity, but otherwise do not differentiate between head and tail items. We find, however, that this explanation does not adequately capture the empirical variation in user behavior. Specifically, relative to this null model, in reality light users disproportionately prefer the head while heavy users disproportionately prefer the tail.

Finally, it has been argued that consumers generally appreciate the tail less than the head, in turn diminishing the importance of large inventories [9, 13]. Although we do find that popular movies receive the highest user ratings, the opposite appears to be true with music: the highest average ratings on Yahoo! Music are given to the most obscure songs. Furthermore, even in the case of movies, typical users regularly give high ratings to tail inventory, suggesting that users not only consume but in fact value specialty items.

One possible objection to these conclusions is that our results are a product of self-selection bias. Netflix users, that is, may be precisely those individuals already frustrated with the limited selection of brick-and-mortar competitors. A related worry is that recommender systems are driving the consumption of specialty items, and thus the long tail does not reflect organic consumer demand. In particular, even if typical individuals are consuming—and appreciating—niche items, they may not miss the absence of such selection. We believe these concerns are mitigated by the consistency of our findings across several diverse domains—movies, music, Web search, and Web browsing. Moreover, in Web search and Web browsing, the effects of selection bias and recommender systems seem minimal.

The Internet, and infinite-inventory retailers in particular, have had a profound and still evolving effect on consumers. The substantial consumption we observe of niche products is likely due to a combination of demand unfulfilled by traditional retailers, decreased search costs for online inventories, and recommender systems that promote specialty items [6]. It remains an important project to further identify and disentangle the root causes and consequences of consumers' taste for the obscure.

Acknowledgments

We thank Mainak Mazumdar and the Nielsen Company for providing Web browsing data, and Prabhakar Raghavan and Duncan Watts for helpful conversations.

6. REFERENCES

[1] Chris Anderson. The long tail. *Wired Magazine*, 12(10):170–177, 2004.

[2] Chris Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, 2006.

[3] Pierre Brémaud. *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 2008.

[4] Erik Brynjolfsson, Yu Jeffrey Hu, and Duncan Simester. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. MIT Center for Digital Business Working Paper, 2007.

[5] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D. Smith. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. MIT Center for Digital Business Working Paper (No. 4305-03), 2003.

[6] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D. Smith. From niches to riches: Anatomy of the long tail. *MIT Sloan Management Review*, 47(4):67–71, 2006.

[7] Doug Downey, Susan Dumais, and Eric Horvitz. Heads and tails: Studies of web search with common and rare queries. In *SIGIR*, 2007.

[8] Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. Understanding the relationship between searchers' queries and information goals. In *CIKM*, 2008.

[9] Anita Elberse. Should you invest in the long tail? *Harvard Business Review*, 86(7/8):88–96, 2008.

[10] Anita Elberse and Felix Oberholzer-Gee. Superstars and underdogs: An examination of the long tail phenomenon in video sales. Harvard Business School Working Paper, No. 07-015, 2006.

[11] Bernard Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36:207–227, 2000.

[12] Lawrence W. Levine. *Highbrow/Lowbrow: The Emergence of Cultural Hierarchy in America*. Harvard University Press, 1988.

[13] William N. McPhee. *Formal Theories of Mass Behavior*. Free Press of Glencoe, 1963.

[14] Richard A. Peterson and Roger M. Kern. Changing highbrow taste: From snob to omnivore. *American Sociological Review*, 61:900–907, 1996.

[15] Richard A. Peterson and Albert Simkus. How musical taste groups mark occupational status groups. In M. Lamont and M. Fournier, editors, *Symbolic Boundaries and the Making of Inequality*, pages 152–68. University of Chicago Press, 1992.

[16] Amanda Spink, Dietmar Wolfram, Bernard Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *JASIST*, 52(3):226–234, 2001.

[17] Tom F. Tan and Serguei Netessine. Is Tom Cruise threatened? Using Netflix Prize data to examine the long tail of electronic commerce. Working Paper, 2009.